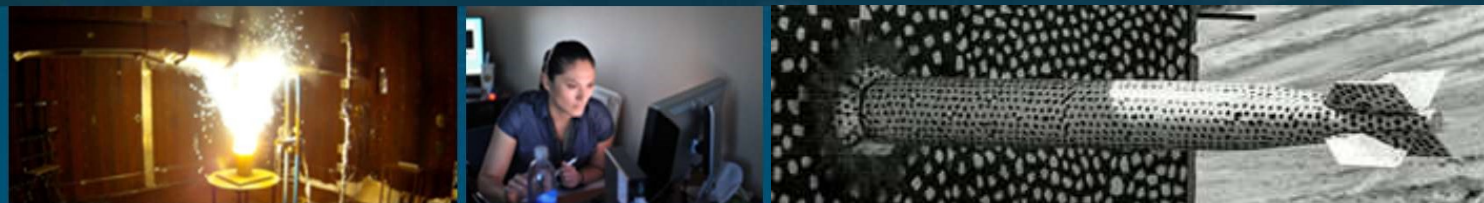


Applied Superconductivity Conference (ASC 2020)
Special Session: Novel Computing - Reversible and Neuromorphic



Reversible Computing as a Path Forward for Improving Dissipation-Delay Efficiency in Superconducting Computing



Tuesday, November 3rd, 2020

Michael P. Frank, Center for Computing Research

with collaborators: Rupert Lewis & Nancy Missert (Sandia), Kevin Osborn & Linqi Yu (LPS), Erik DeBenedictis (Zettaflops, LLC), Karpur Shukla (Brown), Rudro Biswas & Dewan Woods (Purdue), Tom Conte & Anirudh Jain (Georgia Tech).

Approved for public release, SAND2020-11420 C




Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Outline of Talk



Reversible Computing as a Path Forward for Improving Dissipation-Delay Efficiency in Superconducting Computing

- I. Motivation & History
 - The end of cost-efficiency improvements from traditional scaling
 - Rationale for reversible computing from fundamental physics
 - Early implementation concepts
 - Why *dissipation-delay efficiency* is a key performance metric for computing
 - Past attempts to analyze limits on dissipation-delay products (Likharev, Feynman).
- II. Contemporary Technologies for Reversible Computing
 - Reversible adiabatic CMOS
 - Reversible adiabatic superconducting logic
- III. Emerging Ballistic Approaches
 - Brief overview of fluxon-based reversible computing approaches
 - Ballistic Asynchronous Reversible Computing in Superconductors (BARCS) 
- IV. Looking Ahead
 - Fundamental limits of dissipation-delay products from nonequilibrium quantum thermodynamics? (Brown)
 - Impact of hypothetical new reversible technologies at the architecture/systems level? (Georgia Tech.)





Section I. Motivation & History

Reversible Computing as a Path Forward for Improving
Dissipation-Delay Efficiency in Superconducting Computing



Why are we here?

- Progress in the energy-efficiency of the conventional (non-reversible) computing paradigm is approaching hard limits, which ultimately trace back to fundamental thermodynamic issues.
 - Industry is already struggling to continue to advance along the traditional scaling path.
- Energy efficiency is a fundamental limiting factor on the economic utility of computing.
 - Without energy efficiency gains, there are diminishing returns from optimizing *every* other aspect of computing.
- Transitioning to the unconventional computing paradigm known as *reversible computing* provides the only physically possible alternative scaling path for allowing the energy efficiency of *general digital* computing to continue improving indefinitely...
 - And, so far, no fundamental limit to the (even practically) achievable efficiency is known.
- The overall economy is becoming increasingly dependent on computing, as a larger and larger share of economic activity takes place in the cyber realm...
 - Making reversible computing practical thus has the potential to expand *the total future economic value of civilization* (for any given amount of available energy resources) by *indefinitely many* orders of magnitude.

Motivation from Economics / Systems Engineering



In general, *efficiency* η of any process can be defined as the amount P of some valued *product produced* by the process, divided by the amount C of *cost consumed* (in terms of resources, or dollars) by the process.

$$\eta = \frac{P}{C}$$

- For a computing system,
 - P can be amount of useful *information processing performed* (e.g., number of operations) by the system over its operating lifetime, and
 - C can be expressed the sum of manufacturing (& deployment) costs, plus operating costs over the system lifetime.
- We can also annualize the costs, in terms of, e.g. time-amortized manufacturing cost.
 - More sophisticated variations that account for net present value of future returns, depreciation curves, *etc.*, not considered here.
- Operating costs largely amount to *energy-proportioned costs*: $C_{\text{oper}} = c_{\text{en}} \cdot E_{\text{oper}}$
 - c_{en} = operating cost per unit of energy dissipated; E_{oper} = total energy dissipated during a given period of operation.

$$C = C_{\text{tot}} = C_{\text{mfg}} + C_{\text{oper}} \\ \text{(may be time-amortized)}$$

We can thus reduce the efficiency formula $\eta = P/C_{\text{tot}}$ for computing to the form at right:

- E_{op} = Energy dissipated due to *one* primitive device operation (or by one primitive device in time t_d).
- $c_{\text{dev},t}$ = Amortized manufacturing cost per primitive device per unit time t .

$$\eta = \frac{1}{c_{\text{en}} \cdot E_{\text{op}} + c_{\text{dev},t} \cdot t_d} \\ = \frac{1}{E_{\text{op}} t_d \left(\frac{c_{\text{en}}}{t_d} + \frac{c_{\text{dev},t}}{E_{\text{op}}} \right)}$$

Some observations from this equation.:

- There are *diminishing* efficiency returns from decreasing *either* E_{op} or the $c_{\text{dev},t} \cdot t_d$ term in isolation
 - \therefore Continuing to push non-reversible technologies will ultimately reach a dead end!
- Note that if *both* E_{op} and $c_{\text{dev},t}$ were decreased by $N\times$, overall efficiency would be increased by $N\times$. (All else being equal.)
- Decreasing $E_{\text{op}} \cdot t_d$ (dissipation-delay product, DdP) is *often* (but not always!) a win.
 - E.g., in scenarios where total lifetime cost of operation starts out very heavily energy-dominated, total cost can be reduced by lowering E_{op} , *even* in cases where $E_{\text{op}} t_d$ stays the same, or even increases somewhat!
- However, at any given per-device cost, decreasing $E_{\text{op}}(t_d)$ (dissipation as a function of delay) for any given delay value t_d is *always a win*.
 - Thus, this will be our focus in future work.

Semiconductor Roadmap is Ending...

Thermal noise on gate electrodes of minimum-width segments of FET gates leads to significant channel PES fluctuations when $E_g \lesssim 1\text{-}2\text{ eV}$

- Increases leakage, impairs practical device performance
 - Thus, roadmap has minimum gate energy asymptoting to $\sim 2\text{ eV}$

Also, real logic circuits incur many *compounding* overhead factors *multiplying* this limit:

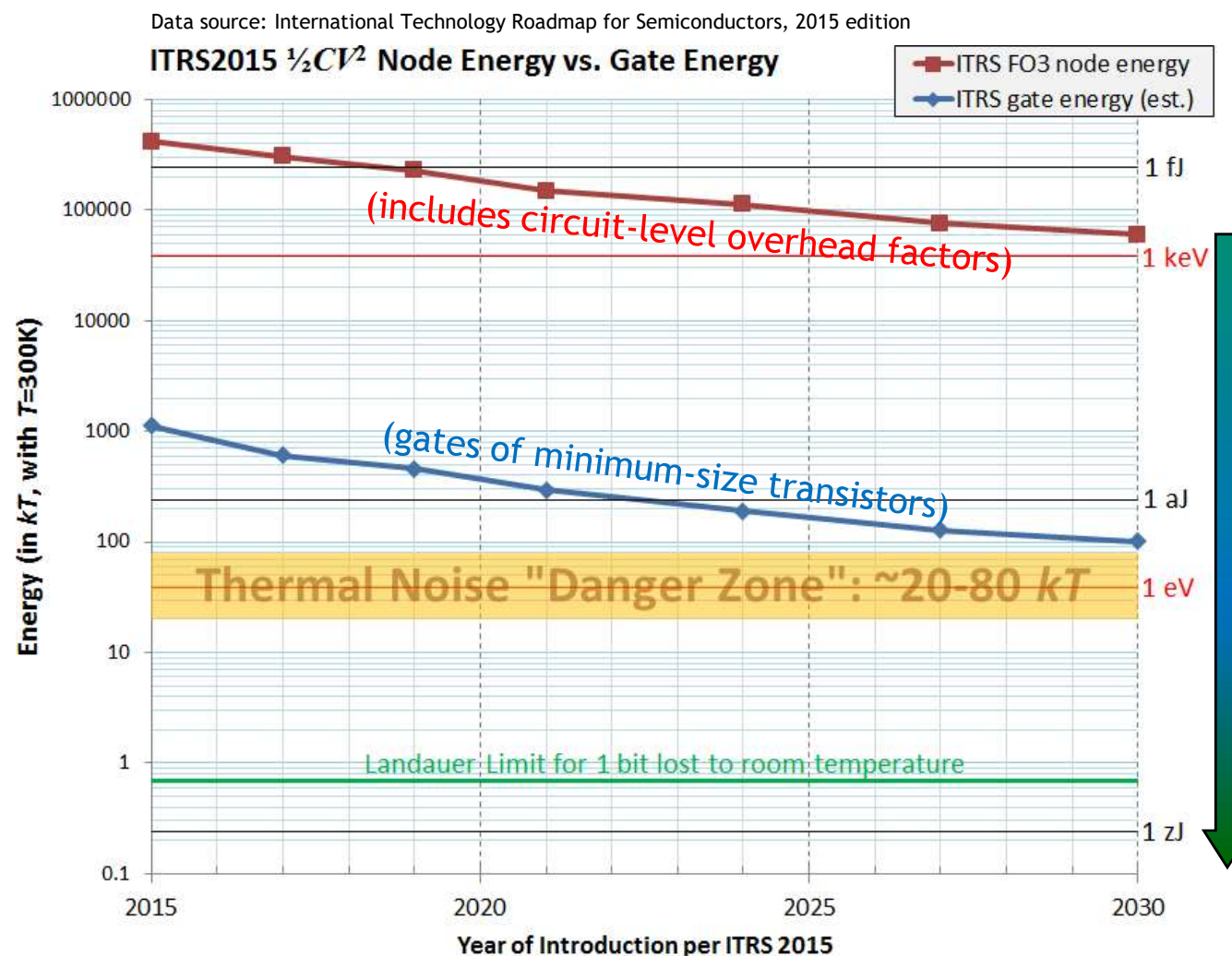
- Transistor width $10\text{-}20\times$ minimum width in fast logic.
- Parasitic (junction, etc.) transistor capacitances ($\sim 2\times$).
- Multiple (~ 2) transistors fed by each input to a given logic gate.
- Fan-out of each gate to a few (~ 3) downstream logic gates.
- Parasitic wire capacitance ($\sim 2\times$).

Due to all these overhead factors, the energy of each logic bit in real logic circuits is many times larger than the minimum-width gate energy!

- $375\text{-}600\times$ (!) larger in ITRS'15.
 - \therefore Practical bit energy for irreversible logic asymptotes to $\sim 1\text{ keV}$!

Practical, real-world logic circuit designs can't just magically cross this $\sim 500\times$ architectural gap!

- \therefore Thermodynamic limits imply much larger practical limits!
 - The end is near!



Only reversible computing can take us from $\sim 1\text{ keV}$ at the end of the CMOS roadmap, all the way down to $\ll kT$.

Reversible computing to the rescue!

The fundamental physical arguments for reversible computing:

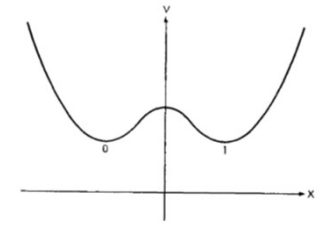
- Landauer's Principle (when properly understood) fundamentally limits the energy efficiency of conventional, non-reversible approaches to general digital computing.
 - Various critics of this statement simply have basic conceptual misunderstandings.
- Physical mechanisms for computing that are *logically* reversible can in principle also approach *physical* reversibility, thereby circumventing *all* limits to the energy efficiency of general digital computing.
 - But, how can we *actually implement* reversible computing in a highly efficient and practical way?
 - This presents a significant challenge for the fields of device physics, device & circuit engineering, and computer engineering.

Some early history of physical implementation concepts:

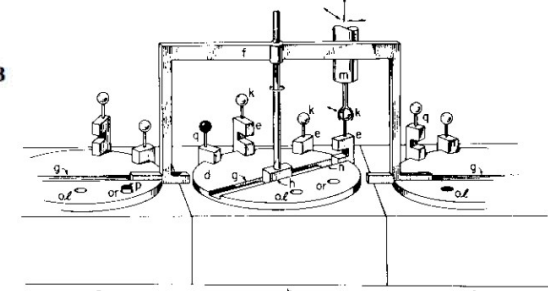
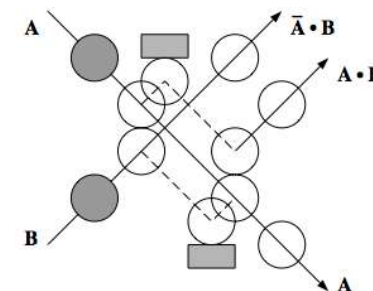
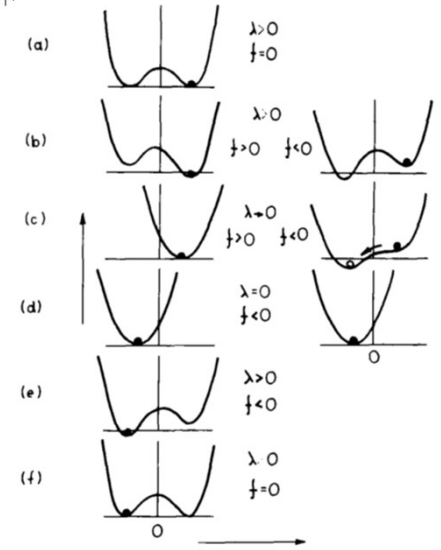
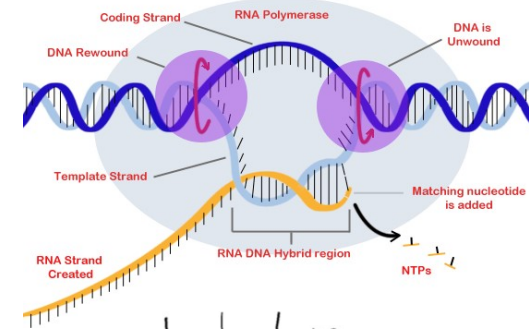
- Landauer (1961)** described physical implementations of reversible computational operations abstractly, in terms of manipulations of **bistable potential energy wells**.
- Bennett (1973)** described logically reversible computations *abstractly* (as Turing machines) and pointed out that biomolecular processes (e.g., DNA transcription) can be understood as computational processes that **operate stochastically** and approach thermodynamic reversibility given appropriate chemical potentials.
- Likharev (1977)** described his Parametric Quantron (PQ) Josephson junction circuit, which could implement reversible transformations of bistable potential energy wells in **superconducting circuits**.
- Fredkin & Toffoli (1980)** described an idealized, **ballistic billiard ball model** (BBM) of reversible computation.
- Bennett (1982)** described a (very slow!) macro-scale **mechanical implementation** of his reversible Turing machine that could operate by Brownian motion.

The rest of this talk will focus on more modern approaches.

- But, many of the same concepts introduced in the early years still apply!



(Landauer '61)



What is dissipation-delay efficiency, and why is it important?

Typically, the *total cost* $\$_{\text{tot}} = \$_E + \$_M$ to perform a computation is minimized when energy-related costs $\$_E$ and manufacturing-related costs $\$_M$ are roughly on the same order.

- Because, there are *diminishing returns* from individually reducing *either one* of these two cost components far below the other one.
 - And, doing so actually makes the total *larger*, if the other cost component gets *increased* as a result.

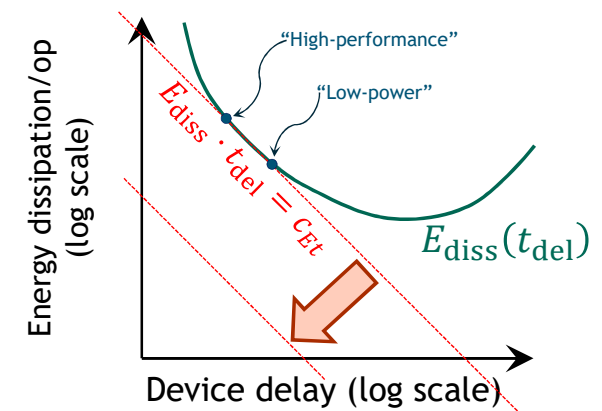
Can express total cost in terms of device parameters: $\$_{\text{tot}} = k_E E_{\text{diss}} + k_M t_{\text{del}}$

For *any* technology that permits tradeoffs between energy efficiency and serial performance, there will be *some* region of the energy-delay curve where the tangent line (on a log-log chart) has slope -1 .

- In this region, the *energy-delay product* is roughly constant.
 - This is even true for voltage scaling in standard irreversible CMOS.
 - But, fully adiabatic techniques can extend this scaling region over a much wider range.
- Different operating points in this linear scaling region will be suitable for applications with different cost *coefficients* k_E, k_M that apply to energy vs. manufacturing cost.
 - E.g., in spacecraft, the effective cost of energy vs. hardware is much greater than in grid-tied applications.

NOTE: If you can move to a new technology whose energy-delay frontier (curve) touches a min. energy-delay product line that is $N\times$ lower than before,

- Then it follows that *total cost* for some applications is reduced by at least $\sqrt{N}\times$!



Dissipation-delay product:

$$C_{Et} = E_{\text{diss}} \cdot t_{\text{del}}$$

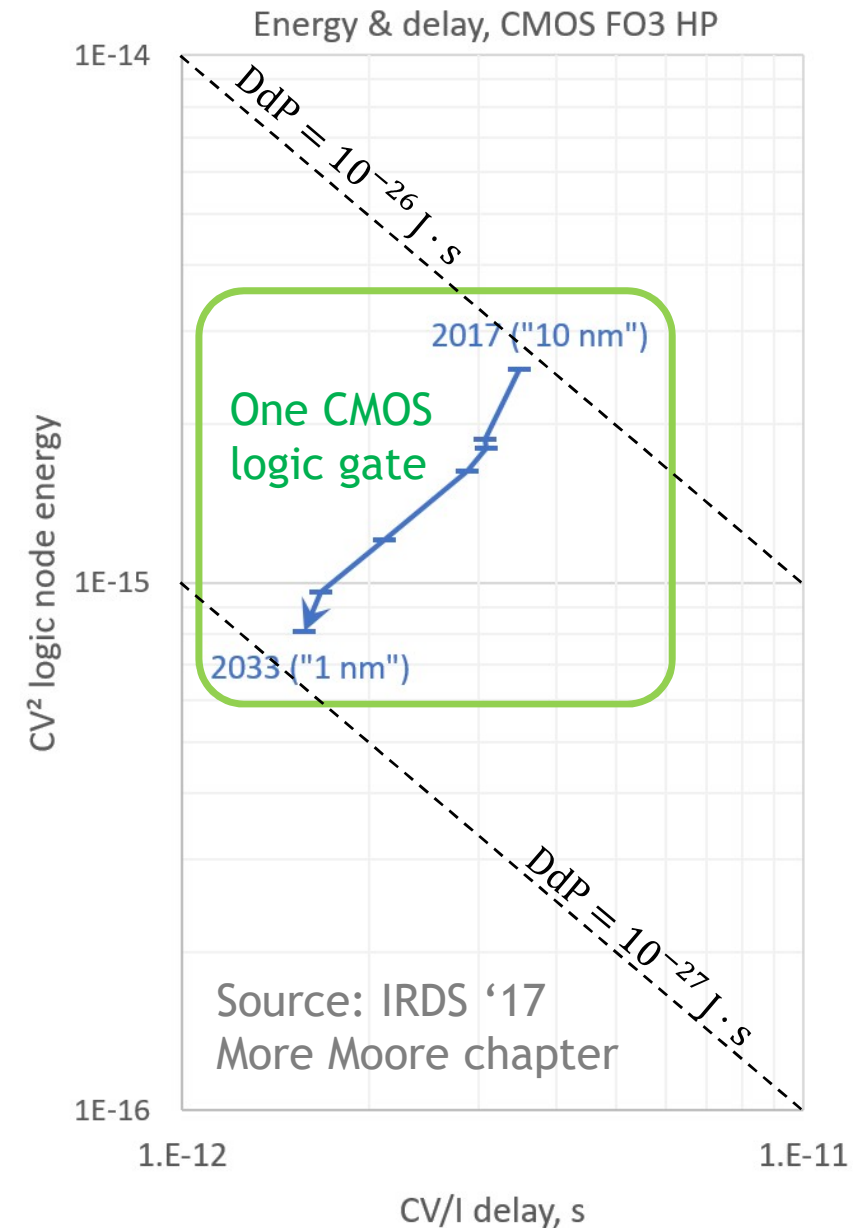
Dissipation-delay efficiency:

$$\eta_{Et} = \frac{1}{C_{Et}}$$

Existing Dissipation-Delay Products (DdP) —Non-reversible Semiconductor Circuits

Conventional (non-reversible) CMOS Technology:

- Recent roadmaps (e.g., IRDS '17) show Dissipation-delay Product (DdP) decreasing by only $< \sim 10\times$ from now to the end of the roadmap (~ 2033).
 - Note the typical dissipation (per logic bit) at end-of-roadmap is projected to be $\sim 0.8 \text{ fJ} = 800 \text{ aJ} = \sim 5,000 \text{ eV}$.
- Optimistically, let's suppose that ways might be found to lower dissipation by an additional $10\times$ beyond even that point.
 - That still puts us at $80 \text{ aJ} = \sim 500 \text{ eV}$ per bit.
- We need at least $\sim 1 \text{ eV} \approx 40 kT$ electrostatic energy at a minimum-sized transistor gate to maintain reasonably low leakage despite thermal noise,
 - And, typical *structural* overhead factors *compounding* this within fast random logic circuits are roughly $500\times$,
 - so, $\sim 500 \text{ eV}$ is *indeed* probably about the practical limit.
 - At least, this is a reasonable order-of-magnitude estimate.





Section II. Contemporary Technologies for Reversible Computing

Device & Circuit Technologies for Reversible Computing—
An Introduction

Adiabatic Charging via MOSFETs

A simple voltage ramp can *approximate* an ideal constant-current source.

- Note that the load gets charged up *conditionally*, if the MOSFET is turned on (gate voltage $V_g \gtrsim V + V_t$) during ramp.
- V_t is the transistor's threshold, typically $< 1/2$ volt

Can discharge the load later using a similar ramp.

- Either through the same path, or a different path.

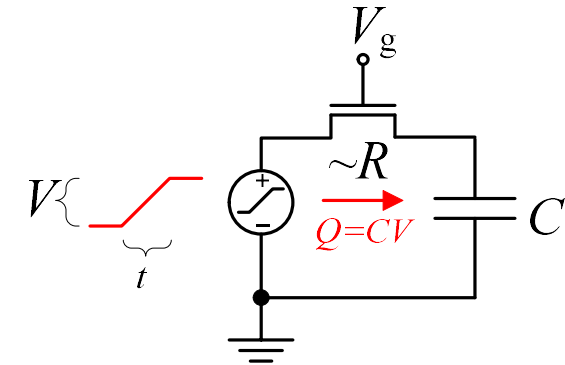
$$t \gg RC \Rightarrow E_{\text{diss}} \rightarrow CV^2 \frac{RC}{t}$$

$$t \ll RC \Rightarrow E_{\text{diss}} \rightarrow \frac{1}{2} CV^2$$

The (ideal) operation of this circuit approaches *physical reversibility* ($E_{\text{diss}} \rightarrow 0$) in the limit $t \rightarrow \infty$, but *only* if a certain *precondition* on the initial state is met (namely, $V_g \gtrsim V_{\text{max}} + V_t$)

- How does the possible physical reversibility of this circuit relate to its *computational* function, and to some *appropriate* concept of logical reversibility?
- Traditional (Landauer/Fredkin/Toffoli) reversible computing theory does not adequately address this question, so, we need a more powerful theory!
- The theory of **Generalized Reversible Computing** (GRC) meets this need.

See [arxiv:1806.10183](https://arxiv.org/abs/1806.10183) for the full GRC model.



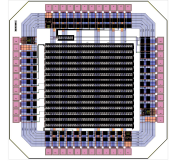
Exact formula for linear ramps:

$$E_{\text{diss}} = s[1 + s(e^{-1/s} - 1)]CV^2$$

given *speed fraction* $s = RC/t$.

Perfectly Adiabatic Reversible Computing in CMOS

2LAL test chip
taped out at
Sandia, Aug. '20



To approach ideal reversible computing in CMOS...

We must aggressively eliminate *all* sources of non-adiabatic dissipation, including:

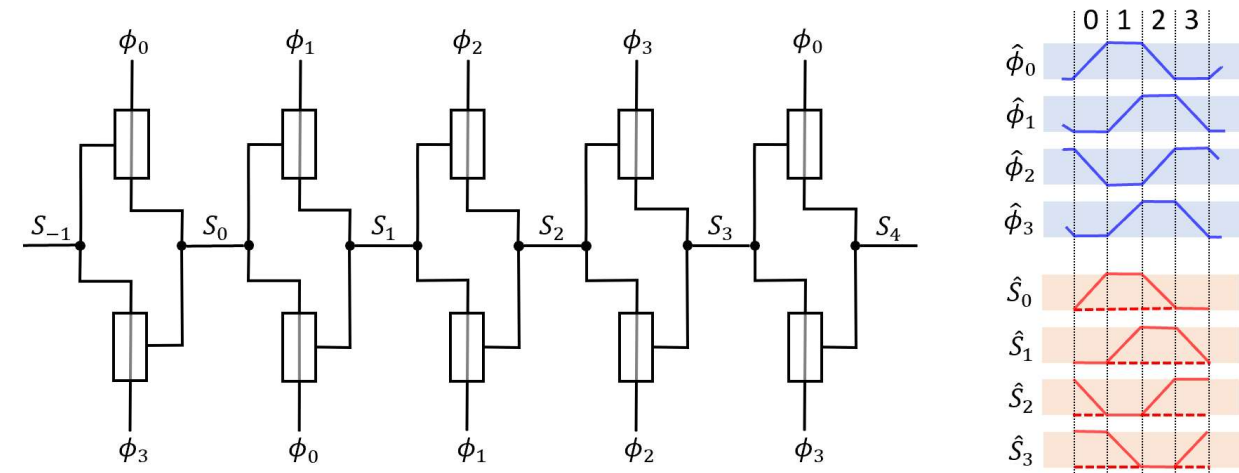
- Diodes in charging path, “sparking,” “squelching,”
 - Eliminated by “**truly, fully adiabatic**” design. (E.g., CRL, 2LAL).
 - Suffices to get to a few aJ (10s of eV) in 180 nm *before* voltage optimization.
- Voltage level mismatches that dynamically arise on floating nodes before reconnection.
 - Eliminated by static, “**perfectly adiabatic**” design. (E.g., S2LAL).

We must also aggressively minimize standby power dissipation from leakage, including:

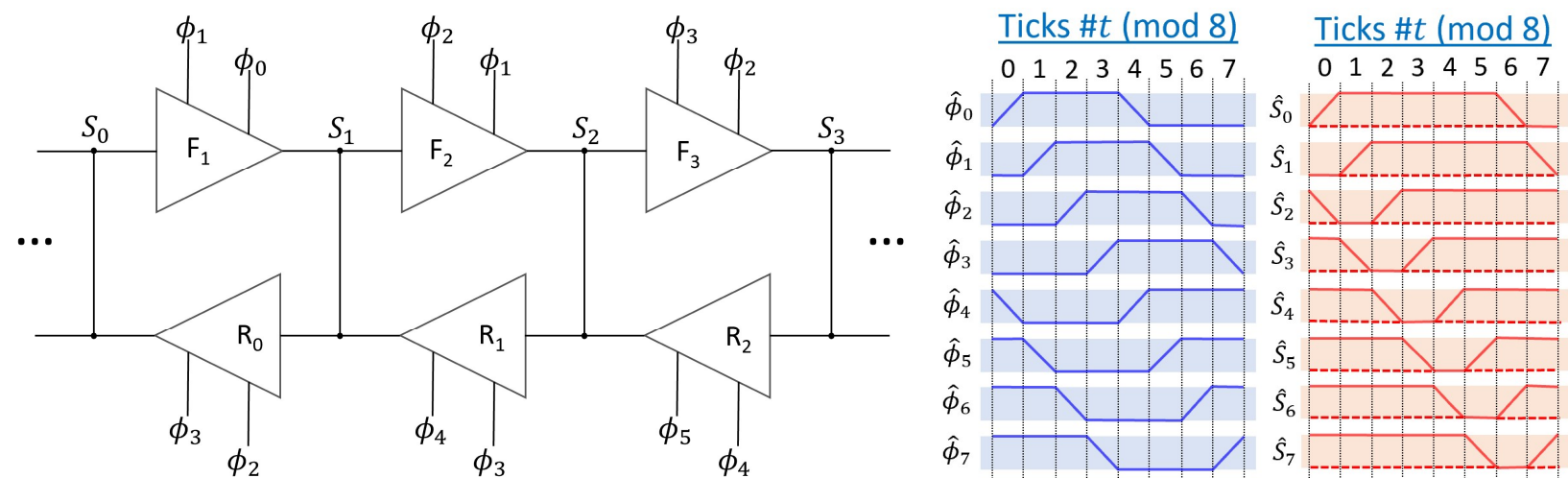
- Subthreshold channel currents
 - Low- T operation helps with this
- Tunneling through gate oxide
 - E.g., use thicker gate oxides

Note: (Conditional) logical reversibility *follows from* perfect adiabaticity.

Shift Register Structure and Timing in 2LAL



Shift Register Structure and Timing in S2LAL



(arxiv:2009.00448)

Adiabatic Reversible Computing in Superconducting Circuits



Work along this general line has roots that go all the way back to Likharev, 1977.

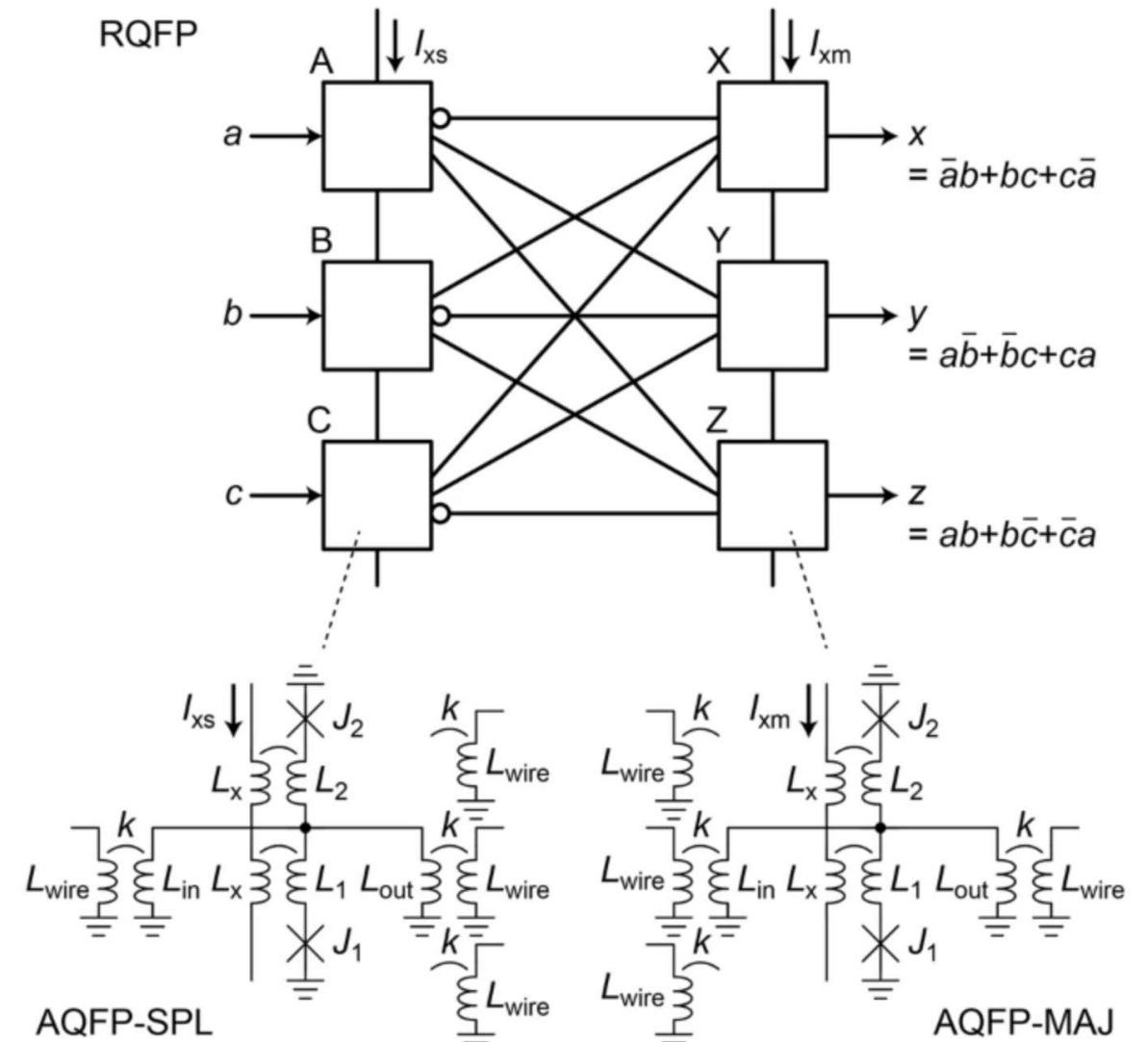
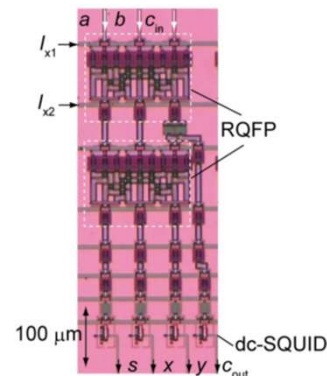
Most active group at present is Prof. Yoshikawa's group at Yokohama National University in Japan.

Logic style called *Reversible Quantum Flux Parametron* (RQFP).

Shown at right is a 3-output *reversible majority gate*.

Full adder circuits have also been built and tested.

Simulations indicate that RQFP circuits can dissipate $< kT \ln 2$ even at $T = 4\text{K}$, at speeds on the order of 10 MHz



Existing Dissipation-Delay Products (DdP)— Adiabatic Reversible Superconducting Circuits

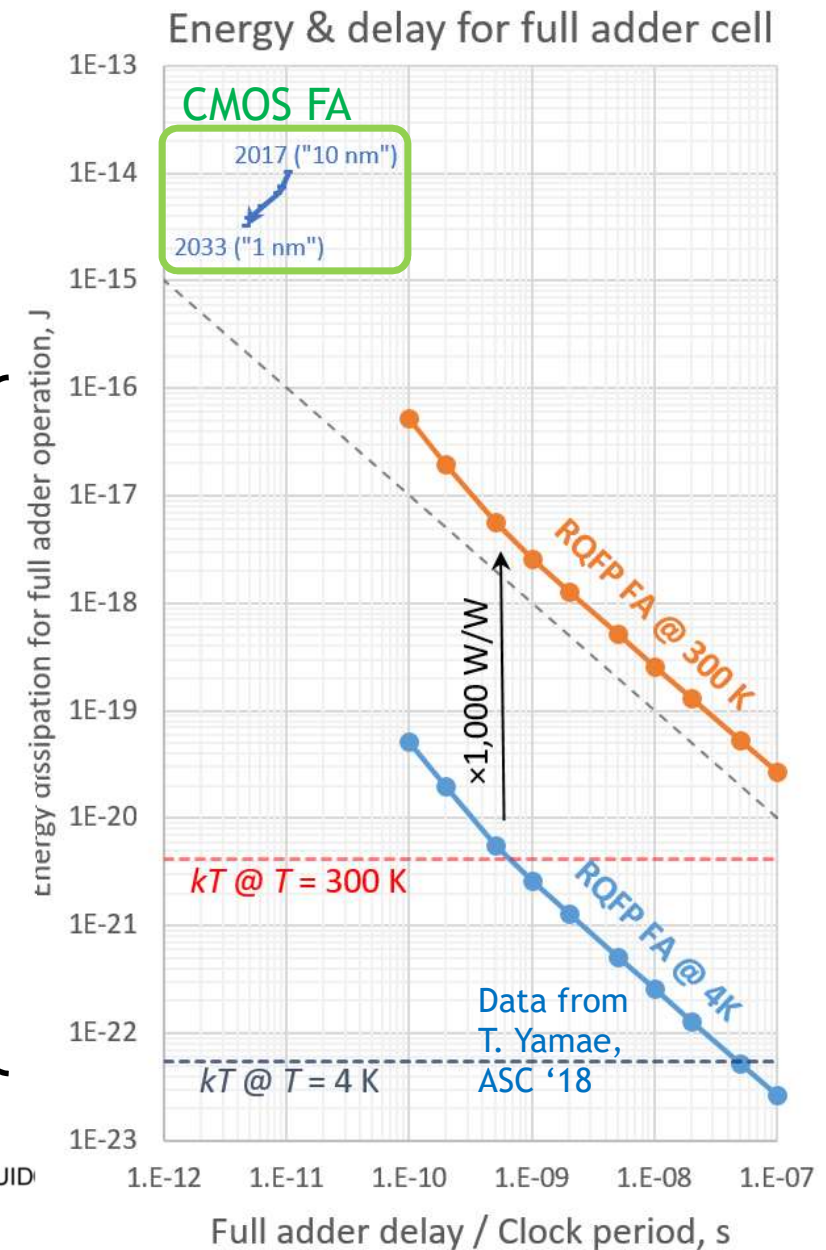
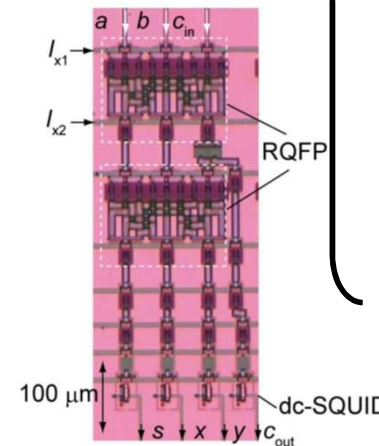
Reversible adiabatic superconductor logic:

- State-of-the-art is the **RQFP** (Reversible Quantum Flux Parametron) technology from Yokohama National University in Japan.
- Chips were fabricated, function validated.
- Circuit simulations predict DdP is $>1,000\times$ *lower* than even *end-of-roadmap* CMOS.
- Dissipation extends *far below* the 300K Landauer limit (and even below the Landauer limit at 4K).
- DdP is *still* better than CMOS even after adjusting by a conservative factor for large-scale cooling overhead (1,000 \times).

Question: Could some *other* reversible technology do even better than this?

- We have a project at Sandia exploring one possible superconductor-based approach for this (more later)...
- But, what are the *fundamental* (technology-independent) limits, if any?

RQFP =
Reversible
Quantum Flux
Parametron
(Yokohama U.)





Section III. Emerging Ballistic Approaches

Reversible Computing as a Path Forward for Improving
Dissipation-Delay Efficiency in Superconducting Computing

Can dissipation scale better than linearly with speed?



Some observations from Pidaparthi & Lent (2018) suggest Yes!

- Landau-Zener (1932) formula for quantum transitions in e.g. scattering processes with a missed level crossing...
 - Probability of exciting the high-energy state (which then decays dissipatively) scales down *exponentially* as a function of speed...

$$P_D = e^{-2\pi\Gamma}$$
 - This scaling is commonly seen in many quantum systems!
- Thus, dissipation-delay *product* may have *no lower bound* for quantum adiabatic transitions—if this kind of scaling can actually be realized in practice.
 - I.e.*, in the context of a complete engineered system.
- Question:** Will unmodeled details (e.g., in the driving system) fundamentally prevent this, or not?

J. Low Power Electron. Appl. 2018, 8(3), 30; <https://doi.org/10.3390/jlpea8030030>

Open Access Article

Exponentially Adiabatic Switching in Quantum-Dot Cellular Automata

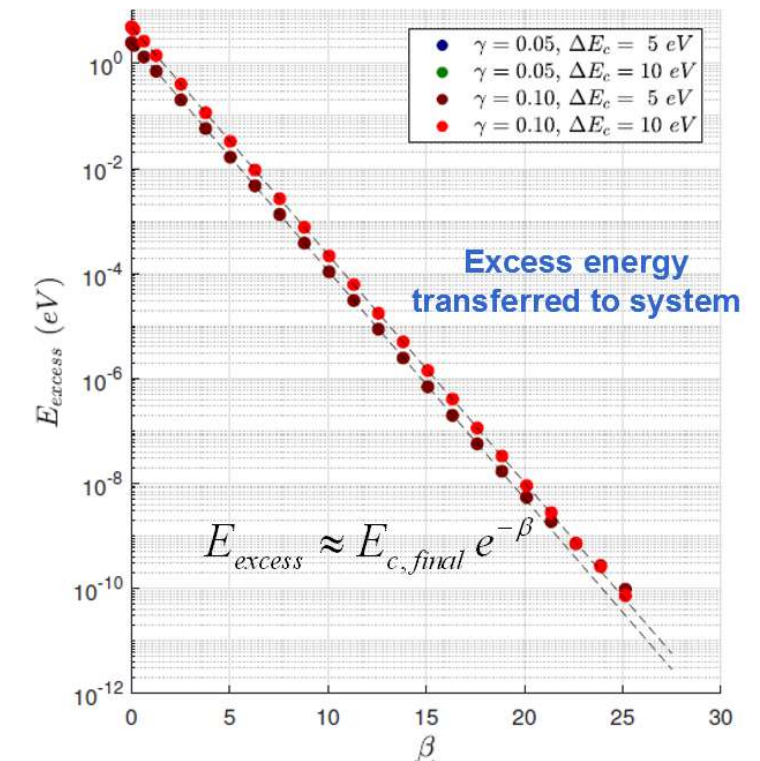
Subhash S. Pidaparthi and Craig S. Lent *

Department of Electrical Engineering, University of Notre Dame, Notre Dame, IN 46556, USA

* Author to whom correspondence should be addressed.

Received: 15 August 2018 / Revised: 5 September 2018 / Accepted: 5 September 2018 / Published: 7 September 2018

(This article belongs to the Special Issue Quantum-Dot Cellular Automata (QCA) and Low Power Application)



Ballistic Reversible Computing

Can we envision reversible computing as a *deterministic* elastic interaction process?

Historical origin of this concept:

- Fredkin & Toffoli's *Billiard Ball Model* of computation ("Conservative Logic," IJTP 1982).
 - Based on elastic collisions between moving objects.
 - Spawned a subfield of "collision-based computing."
 - Using localized pulses/solitons in various media.

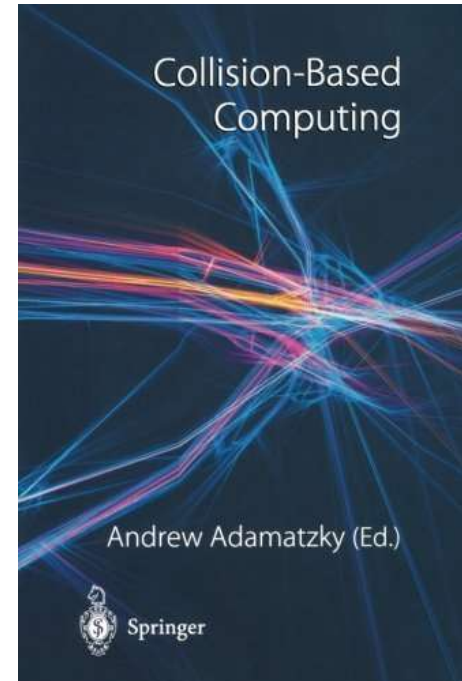
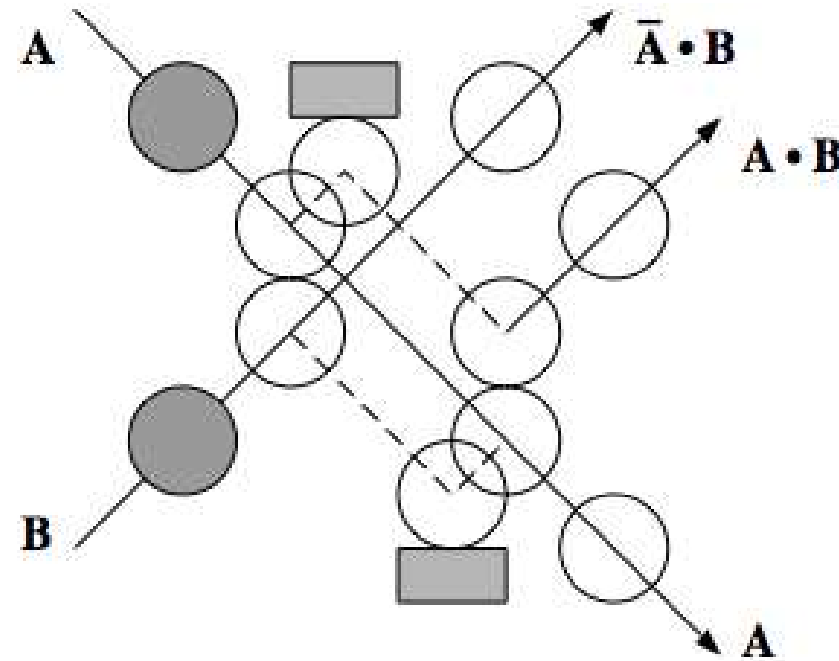
No power-clock driving signals needed!

- Devices operate when data signals arrive.
- The operation energy is carried by the signal itself.
 - Most of the signal energy is preserved in outgoing signals.

However, all (or almost all) of the existing design concepts for ballistic computing invoke implicitly *synchronized* arrivals of ballistically-propagating signals...

- Making this work in reality presents some serious difficulties, however:
 - Unrealistic in practice to assume precise alignment of signal arrival times.
 - Thermal fluctuations & quantum uncertainty, at minimum, are always present.
 - Any relative timing uncertainty leads to chaotic dynamics when signals interact.
 - Exponentially-increasing uncertainties in the dynamical trajectory.
 - Deliberate *resynchronization* of signals whose timing relationship is uncertain incurs an inevitable energy cost.

Can we come up with a new ballistic model that avoids these problems?



Ballistic Asynchronous Reversible Computing (BARC)



Problem: Conservative (dissipationless) dynamical systems generally tend to exhibit chaotic behavior...

- This results from direct nonlinear *interactions* between multiple continuous dynamical degrees of freedom (DOFs), which amplify uncertainties, exponentially compounding them over time...
- E.g., positions/velocities of ballistically-propagating “balls”
 - Or more generally, any localized, cohesive, momentum-bearing entity: Particles, pulses, quasiparticles, solitons...

Core insight: In principle, we can greatly reduce or eliminate this tendency towards dynamical chaos...

- We can do this simply by *avoiding* any direct interaction between continuous DOFs of different ballistically-propagating entities

Require localized pulses to arrive *asynchronously*—and furthermore, at clearly distinct, *non-overlapping* times

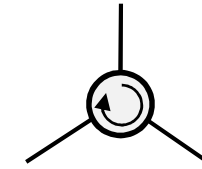
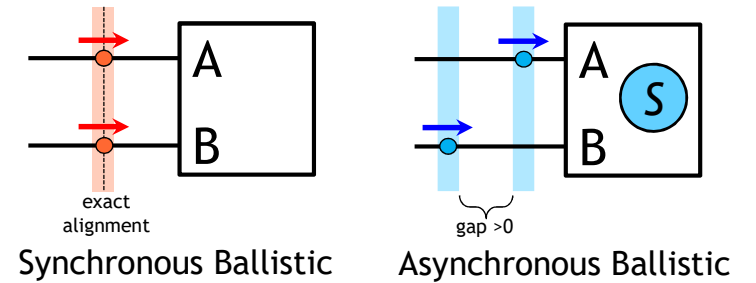
- Device’s dynamical trajectory then becomes *independent* of the precise (absolute *and* relative) pulse arrival times
 - As a result, timing uncertainty per logic stage can now accumulate only *linearly*, not exponentially!
 - Only relatively occasional re-synchronization will be needed
- For devices to still be capable of doing logic, they must now maintain an internal discrete (digitally-precise) state variable—a stable (or at least metastable) stationary state, e.g., a ground state of a well

No power-clock signals, unlike in adiabatic designs!

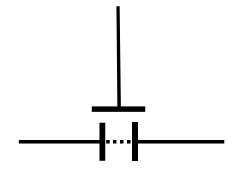
- Devices simply operate whenever data pulses arrive
- The operation energy is carried by the pulse itself
 - Most of the energy is preserved in outgoing pulses
 - Signal restoration can be carried out incrementally

Goal of current effort at Sandia: Demonstrate BARC principles in an implementation based on fluxon dynamics in Superconducting electronics (SCE)

(BARCS effort)

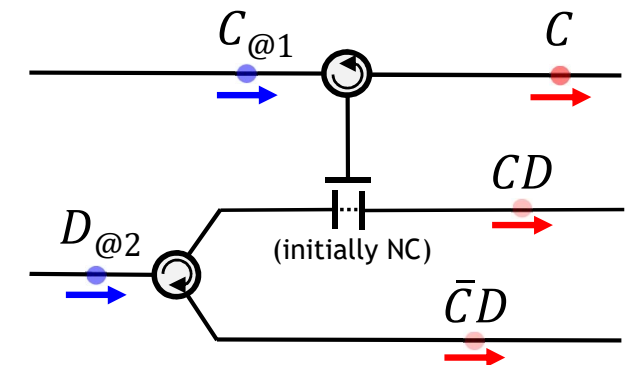


Rotary
(Circulator)



Toggled
Barrier

Example BARC device functions



Example logic construction

Simplest Fluxon-Based (bipolarized) BARC Function



One of our early tasks: Characterize the simplest nontrivial BARC device functionalities, given a few simple design constraints applying to an SCE-based implementation, such as:

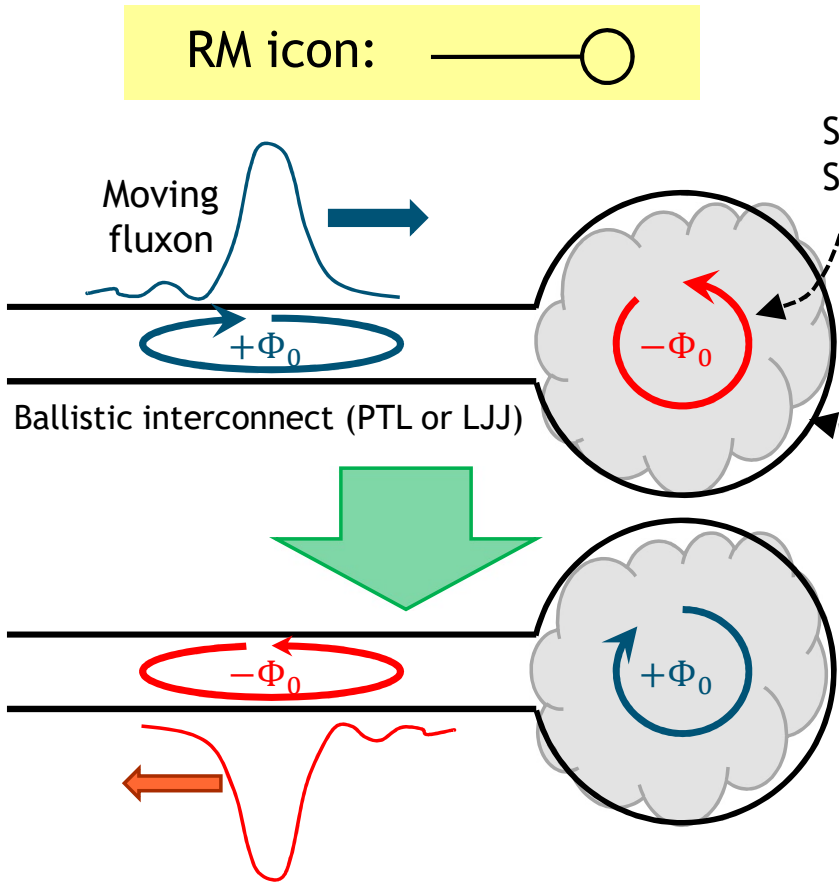
- (1) Bits encoded in fluxon polarity; (2) Bounded planar circuit conserving flux; (3) Physical symmetry.

Determined through theoretical hand-analysis that the simplest such function is the 1-Bit, 1-Port Reversible Memory Cell (RM):

- Due to its simplicity, this was then the preferred target for our subsequent detailed circuit design efforts...

RM Transition Table

Input Syndrome		Output Syndrome
+1(+1)	→	(+1)+1
+1(-1)	→	(+1)-1
-1(+1)	→	(-1)+1
-1(-1)	→	(-1)-1



Some planar, unbiased, reactive SCE circuit w. a continuous superconducting boundary

- Only contains L's, M's, C's, and *unshunted* JJs
- Junctions should mostly be *subcritical* (avoids R_N)
- Conserves total flux, approximately nondissipative

Desired circuit behavior (NOTE: conserves flux, respects T symmetry & logical reversibility):

- If polarities are opposite, they are swapped (shown)
- If polarities are identical, input fluxon reflects back out with no change in polarity (not shown)
- (*Deterministic*) *elastic 'scattering'* type interaction: Input fluxon kinetic energy is (nearly) preserved in output fluxon

RM—First working (in simulation) implementation!

Erik DeBenedictis: “Try just strapping a JJ across that loop.”

- This actually works!

“Entrance” JJ sized to = about 5 LJJ unit cells ($\sim 1/2$ pulse width)

- I first tried it twice as large, & the fluxons annihilated instead...
 - “If a $15\ \mu\text{A}$ JJ rotates by 2π , maybe $1/2$ that will rotate by 4π ” 🤔

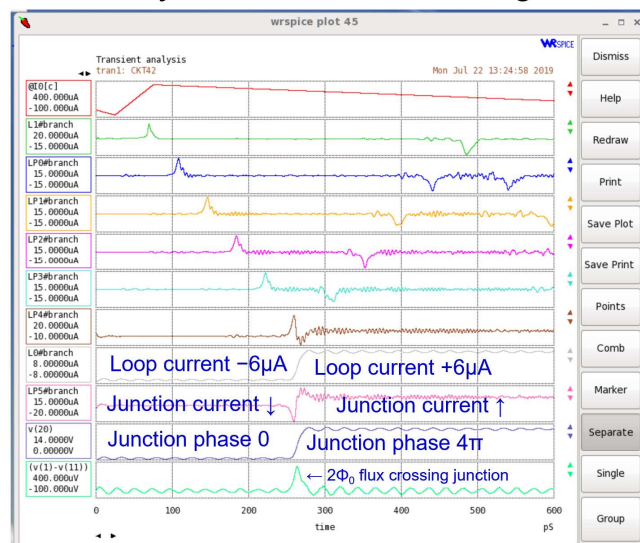
Loop inductor sized so ± 1 SFQ will fit in the loop (but not ± 2)

- JJ is sitting a bit below critical with ± 1

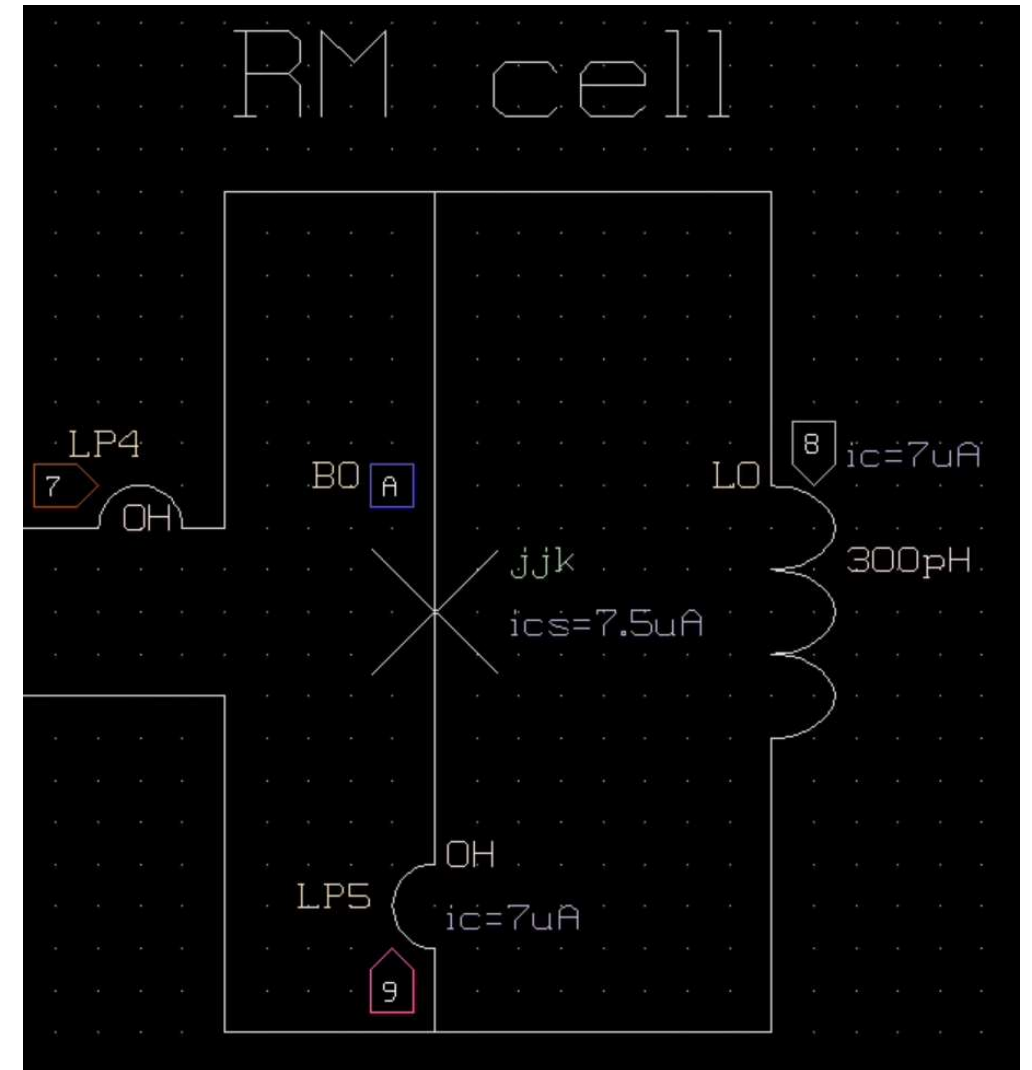
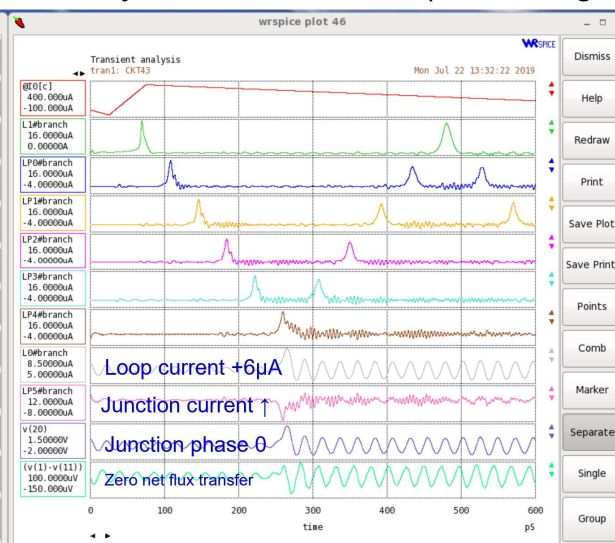
WRspice simulations with ± 1 fluxon initially in the loop

- Uses `ic` parameter, & `uic` option to `.tran` command
 - Produces initial ringing due to overly-constricted initial flux
 - Can damp w. small shunt G

Polarity mismatch \rightarrow Exchange



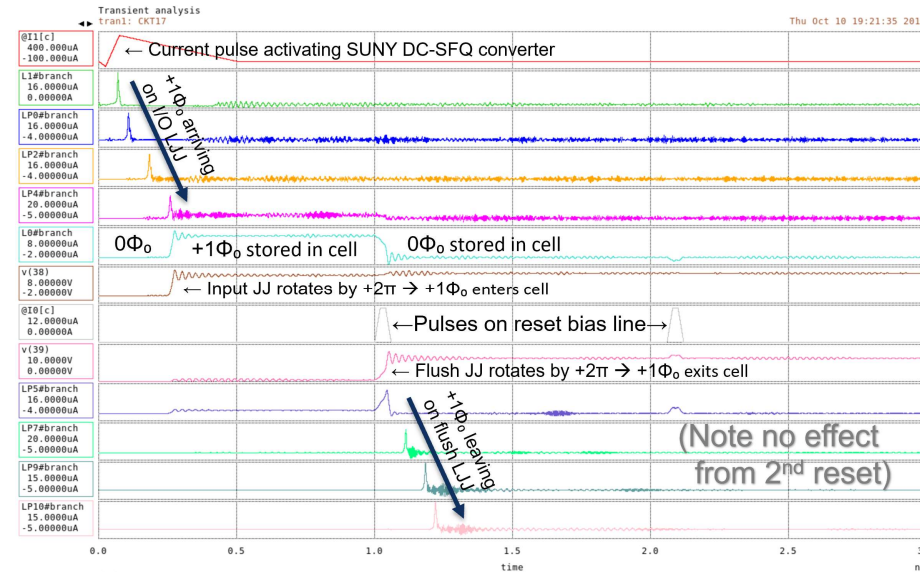
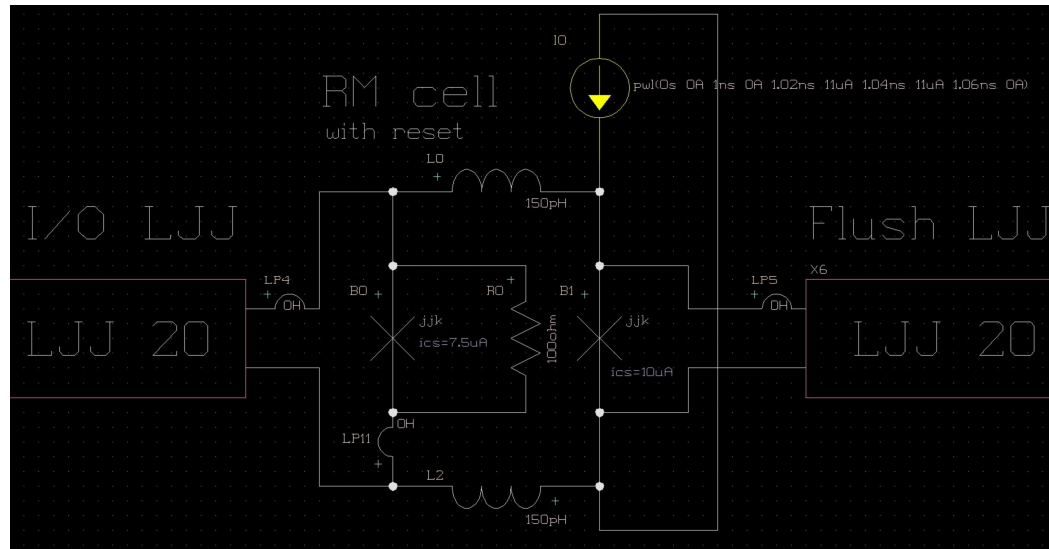
Polarity match \rightarrow Reflect (=Exchange)



Resettable version of RM cell—Designed & Fabricated!

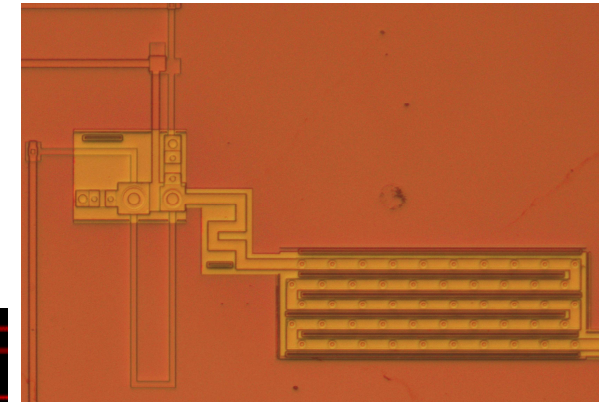
Apply current pulse of appropriate sign to flush the stored flux (the pulse here flushes out positive flux)

- To flush either polarity → Do both (\pm) resets in succession

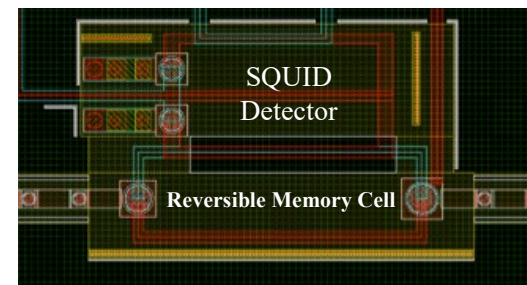
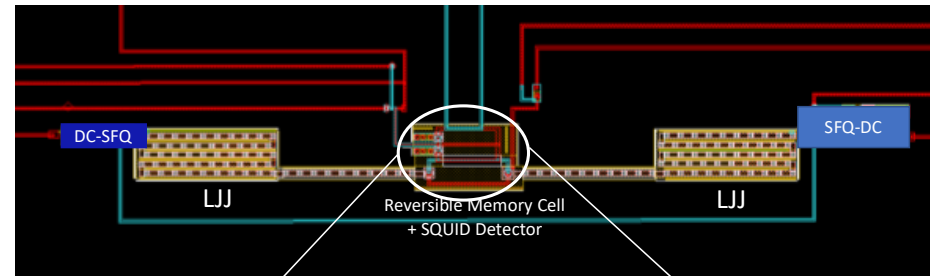
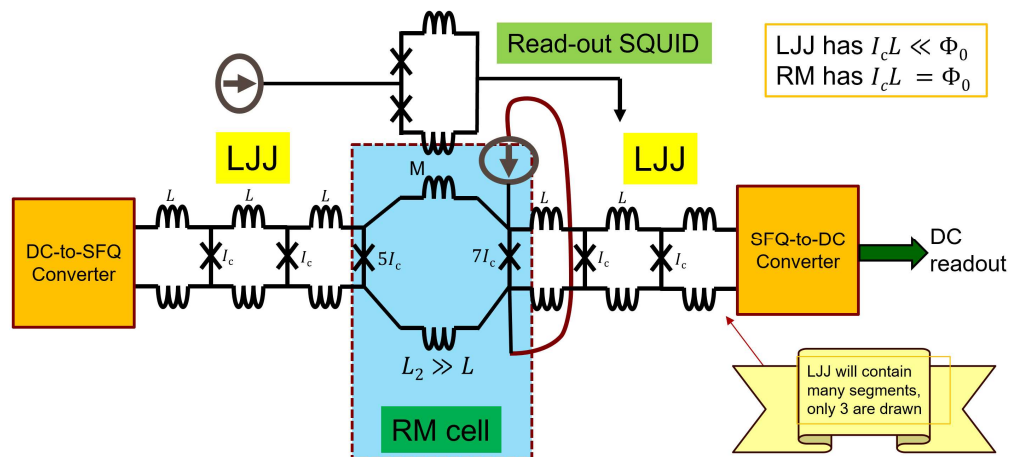
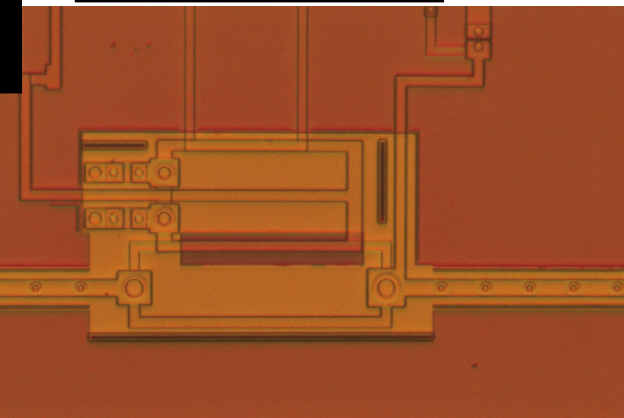


Fabrication at SeeQC with support from ACI

DC-SFQ & LJJ



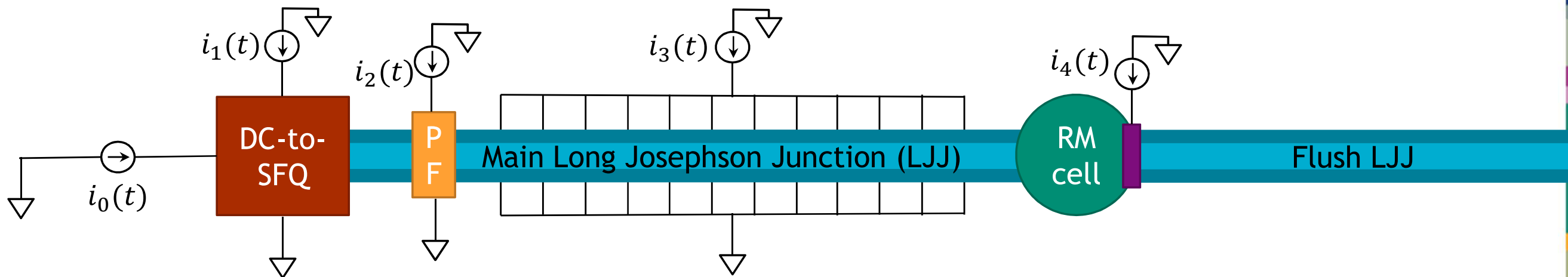
RM Cell & SQUID



Concept for energy dissipation measurements

Experimental protocol:

- Pulse the “flush/reset” JJ using i_4 bias \pm to ensure there is 0 trapped flux in the RM cell initially.
- Use the DC-SFQ converter to inject a $+\Phi_0$ pulse through the polarity filter (PF) into RM cell and store it.
- Use the DC-SFQ converter to inject a $-\Phi_0$ pulse through the polarity filter (PF), and then immediately...
- Turn off the polarity filter (PF)—that is, reset it to 0 bias current (and ideally, tristate it).
- Initiate a *periodic* \pm current bias waveform (symmetric square wave) on the LJJ (i_3).
 - Purpose of this: Alternate between accelerating $-\Phi_0$ pulses to the right and $+\Phi_0$ pulses to the left, vs. the opposite (after reflection off PF).
- At appropriate combinations of amplitude I & frequency f , the i_3 drive signal will hit a resonance.
 - Detect resonance by measuring reflected i_3 power with an RF network analyzer—at resonance there will be a dip in reflected power.
- From the resonance point I, f and the measured S_{11} , we can immediately calculate the following parameters:
 - Fluxon velocity
 - Total energy dissipation per cycle
- To infer what part of the energy dissipation is due to the RM cell:
 - Just do a similar test with a simple inverting reflector (open circuit) in place of the RM cell.



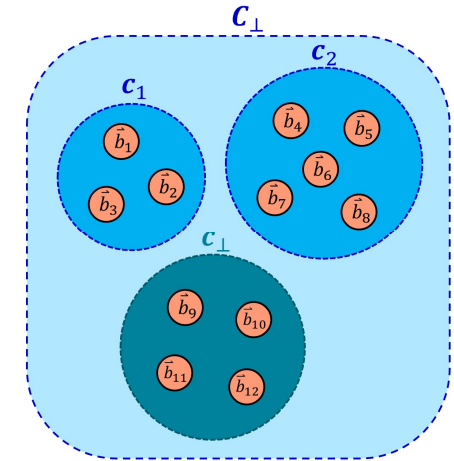


Section IV. Looking Ahead

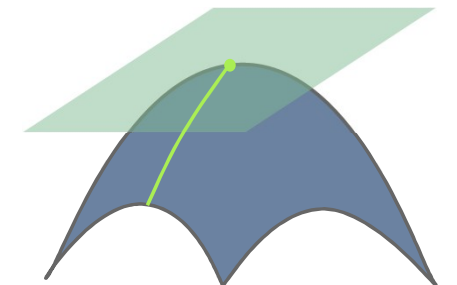
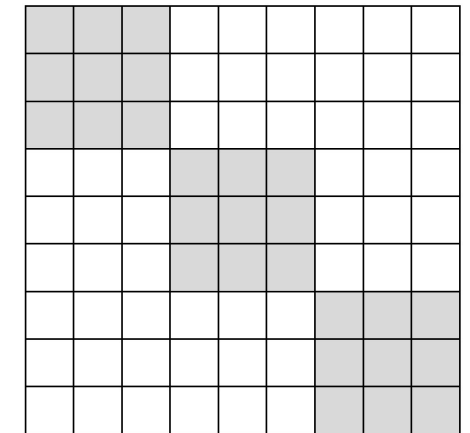
Reversible Computing as a Path Forward for Improving
Dissipation-Delay Efficiency in Superconducting Computing

Fundamental Physics of Reversible Computing

(Work with Karpur Shukla, Brown University)



$$U_s^t(\mathcal{S}, \mathcal{B}) \Vdash \mathcal{C}_s^t(O_s^t, \rho_s)$$



- Goals of this effort:
 - Look for fundamental physical limits of reversible computing
 - *E.g.*, minimum entropy production per operation as a function of delay, temperature, etc.
 - Identify ways to harness exotic quantum phenomena if needed to saturate the limits
- Steps completed so far:
 - Identification of classical computational states with disjoint sets of orthonormal basis states in a (time-dependent, in general) *protocomputational basis* \mathcal{B} .
 - Formalization of what it *means* for a unitary *quantum* evolution U_s^t on a computational system \mathcal{S} (physical computer) to *implement* a given *classical* (and possibly reversible and/or stochastic) computational operation O_s^t between times s and t .
- Research strategy looking forward:
 - Computational states correspond to *decoherence-free subspace blocks* of overall Hilbert space.
 - Quantum Markov equation with multiple asymptotic states: admits subspace dynamics for open systems under Markov evolution.
 - Induces geometric tensor for *manifold of asymptotic states*.
 - Similar to quantum geometric tensor / Berry curvature for closed systems.
 - Current work: use multiple asymptotic state framework to derive thermodynamic quantities...
 - Uncertainty relations, dissipation and dissipation-delay product.

Assessment of Architectural Implications

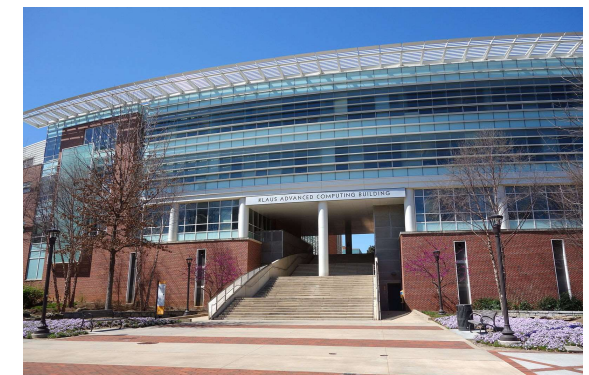
(Work with Tom Conte and Anirudh Jain, Georgia Tech)

Suppose the study of fundamental limits will be successful, and yield a better understanding of the limiting tradeoffs between dissipation, speed, *etc.*

- **Question:** What would be the architectural implications of attaining those limits?
- **Note:** We can begin exploring this question even before the main study yields results!

Research plan for Sandia/GT collaboration:

- **Sandia** defines a common *generic* model of abstract reversible device technologies (including adiabatic and/or asynchronous variants), characterized by key parameters and their scaling, *e.g.*, $E_{\text{diss}}(t_d)$, P_{leak} , *etc.*
- **Georgia Tech** designs a hierarchy of architectural components *composed* out of these generic reversible elements, leading towards a RISC style CPU architecture, including:
 - Multiplexers (32 bits wide, 2-to-1 and 4-to-1).
 - Comparators and Adders (32-bit-wide).
 - Integer Multipliers (32×32 bits, used for address arithmetic).
 - 32-bit ALU (Arithmetic-Logic Unit).
 - Canonical 5-stage pipelined RISC style processor including control unit.
- Meanwhile, **Sandia** supplies various special cases of the generic model reflecting interesting candidate (including hypothetical or preliminary) scaling relations emerging from main study.
 - **Georgia Tech** analyzes the effect of these particular model cases on the efficiency of architectural components
- **Georgia Tech** concludes by:
 - Conducting a study of the pareto optimal frontier of efficiency for *partially*-reversible architectures



Conclusion



The continued advancement of reversible computing technology is a *prerequisite* to prevent the overall cost-efficiency of general digital computing from plateauing in the foreseeable future.

- The ultimate limit of efficiency achievable along this path is still unknown.

Dissipation-delay product (DdP) is a key figure of merit for computing.

- Superconductor technologies may be able to outperform CMOS on this metric.

Existing *adiabatic* approaches may have fundamental limits on their DdP...

- Could we do better using ballistic approaches?
 - A current project at Sandia is investigating this.

There is lots of very interesting/promising work still to be done along these lines!

- We encourage more researchers to begin investigating these approaches.



Summary Slides (for Q&A)

Reversible Computing as a Path Forward for Improving Dissipation-Delay
Efficiency in Superconducting Computing

Reversible Computing as a Path Forward for Improving Dissipation-Delay Efficiency in Superconducting Computing



A few key points:

- Reducing *both* dissipation/op and delay (or device cost) by $N\times$ improves overall system cost-efficiency by $N\times$ (all else being equal).
 - Whereas, not reducing both terms together \rightarrow diminishing returns.
- After the end of the CMOS roadmap in early 2030s, thermal noise issues will prevent much further improvement in energy efficiency for conventional logic.
 - The *only* way to obtain extreme further improvements in energy efficiency (and overall cost-efficiency) for general digital computing will be through reversible computing.
- Reversible adiabatic superconducting logic already has better energy-delay product than even end-of-roadmap CMOS.
 - Adiabatic approaches to reversible computing allow you to trade off dissipation and delay over a wide range, but don't directly allow you to *continue* improving energy-delay product.
- Can we do better using ballistic reversible approaches?
 - Still an open question, not yet ruled out by any existing analyses.
- The Ballistic Asynchronous Reversible Computing (BARC) model was shown computation universal in 2017.
 - In a present project at Sandia, we are working to implement this model in superconducting circuits.
- The first working BARC element (reversible memory cell) was invented in 2019,
 - It was fabricated (and patent filed) this year. Testing is about to start.

